REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
17-02-2016	Final Report	15-Mar-2011 - 14-Mar-2015

4. TITLE AND SUBTITLE
Final Report: Network Data: Statistical Theory and New Models

5a. CONTRACT NUMBER
W911NF-11-1-0114

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Bin Yu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Berkeley
Sponsored Projects Office
2150 Shattuck Avenue, Suite 300
Berkeley, CA 94704 -5940

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
57892-MA.45

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
During this period of review, Bin Yu worked on many thrusts of high-dimensional statistical theory and methodologies. Her research covered a wide range of topics in statistics including analysis and methods for spectral clustering for sparse and structured networks [2,7,8,21], sparse modeling (e.g. Lasso) [4,10,11,17,18,19], statistical guarantees for the EM algorithm [3], statistical analysis of algorithm leveraging for solving big data problems [5], causal network modeling [15,20], stability as a general concept/framework for reproducible statistical discovery [9,12], and high-dimensional inference [12]. Yu also collaborated with other research groups and Labs to conduct

15. SUBJECT TERMS
causal inference, graphical models, stability, cross validation, encoding and decoding of fMRI signals, bootstrapping, Lasso+OLS, confidence interval, concise comparative summarization, EM algorithm, spectral clustering, aerosol retrieval, algorithmic leveraging

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bin Yu
UU	UU	UU	UU		19b. TELEPHONE NUMBER 510-642-2021

Report Title

Final Report: Network Data: Statistical Theory and New Models

ABSTRACT

During this period of review, Bin Yu worked on many thrusts of high-dimensional statistical theory and methodologies. Her research covered a wide range of topics in statistics including analysis and methods for spectral clustering for sparse and structured networks [2,7,8,21], sparse modeling (e.g. Lasso) [4,10,11,17,18,19], statistical guarantees for the EM algorithm [3], statistical analysis of algorithm leveraging for solving big data problems [5], causal network modeling [15,20], stability as a general concept/framework for reproducible statistical discovery [9,13], and high-dimensional inference [12]. Yu also collaborated with other research groups and Labs to conduct interdisciplinary research in areas including systems biology, neuroscience, remote sensing, document summarization, and social networks. For example, she has been collaborating with Dr. Frise et al. on constructing gene-gene interaction networks [1], with the Gallant Lab on understanding visual pathway of primates by using sparse coding [14], and with environmental scientists at JPL and Emory University to retrieval from NASA MISR remote sensing images aerosol index AOD for air pollution monitoring and management [6,16].

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received	Paper
02/08/2016 20.00	Bin Yu, Hanzhong Liu. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression, Electronic Journal of Statistics, (12 2013): 3124. doi: 10.1214/14-EJS875
02/16/2016 29.00	Karl Rohe, Sourav Chatterjee, Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel, The Annals of Statistics, (08 2011): 0. doi: 10.1214/11-AOS887
02/16/2016 43.00	Ping Ma, Michael W. Mahoney, Bin Yu. A Statistical Perspective on Algorithmic Leveraging, Journal of Machine Learning Research, (04 2015): 861. doi:
02/16/2016 42.00	Taesup Moon, Yueqing Wang, Yang Liu, Bin Yu. Evaluation of a MISR-Based High-Resolution Aerosol Retrieval Method Using AERONET DRAGON Campaign Data, IEEE Transactions on Geoscience and Remote Sensing, (08 2015): 0. doi: 10.1109/TGRS.2015.2395722
02/16/2016 41.00	Geoffrey Schiebinger, Martin J. Wainwright, Bin Yu. The geometry of kernelized spectral clustering, The Annals of Statistics, (04 2015): 0. doi: 10.1214/14-AOS1283
02/16/2016 40.00	Chinghway Lim, Bin Yu. Estimation Stability with Cross Validation (ESCV), Journal of Computational and Graphical Statistics, (04 2015): 0. doi: 10.1080/10618600.2015.1020159
02/16/2016 39.00	Garvesh Raskutti, Martin J. Wainwright, Bin Yu. Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule, Journal of Machine Learning Research, (01 2014): 335. doi:
02/16/2016 38.00	Bin Yu. Stability (Invited paper), Bernoulli, (09 2013): 0. doi: 10.3150/13-BEJSP14
02/16/2016 37.00	Garvesh Raskutti, Martin J.Wainwright, Bin Yu. Minimax-Optimal Rates For Sparse Additive Models Over KernelClasses Via Convex Programming, Journal of Machine Learning Research, (02 2012): 389. doi:
02/16/2016 36.00	Yuval Benjamini, Bin Yu. The shuffle estimator for explainable variance in FMRI experiments, The Annals of Applied Statistics, (12 2013): 0. doi: 10.1214/13-AOAS681
02/16/2016 35.00	Jinzhu Jia, Bin Yu, Yangbo He. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs, The Annals of Statistics, (08 2013): 0. doi: 10.1214/13-AOS1125
02/16/2016 34.00	Yueqing Wang, Xin Jiang, Bin Yu, Ming Jiang. A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data, Journal of the American Statistical Association, (06 2013): 0. doi: 10.1080/01621459.2013.796834
02/16/2016 33.00	Jinzhu Jia, Karl Rohe, Bin Yu. The Lasso under Poisson-like heteroscedasticity, Statistica Sinica, (12 2013): 99. doi: 10.5705/ss.2010.254
02/16/2016 30.00	Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, Bin Yu. Geometry of the faithfulness assumption in causal inference, The Annals of Statistics, (04 2013): 0. doi: 10.1214/12-AOS1080

11/06/2014 21.00 Jinzhu Jia, Luke Miratrix, Bin Yu, Brian Gawalt, Laurent El Ghaoui, Luke Barnesmoore, Sophie Clavier. Concise comparative summaries (CCS) of large text corpora with a human experiment, The Annals of Applied Statistics, (03 2014): 0. doi: 10.1214/13-AOAS698

TOTAL: **15**

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

A. Plenary/Main talks
Introductory Overview Talk on Big Data, JSM, Montreal, Aug., 2013
11th International Workshop on Multiple Classifer Systems (MCS 2013), Nanjing University, China, May, 2013
Fong Symposium, SF State University, Mathematics Department, April, 2013

B. Short Courses
Organizer and co-lecturer, Summer School on Applied Statistics and Machine Learning, IMA, June, 2013

C. Invited Presentations
School of Mathematical Sciences, Peking University, Aug., 2014
IMS Presidential Address, IMS Annual Meeting, Sydney, July, 2014
Bioinformatics Div. Seminar, Walter & Eliza Hall Institute of Medical Research, Melbourne, June, 2014
Conference on Big Data, Statistics, UW-Madison, June, 2014
Biostatistics Seminar, Univ of Copenhagen, May, 2014
Abel Symposium, Lofoten, Norway, May, 2014
Workshop on Big Data at Rutgers Univ, May, 2014
Seminar at HITS, Heidelberg, Germany, March, 2014
Oberwolfach Workshop on Adaptive Nonparametric Estimation, Germany, March 2014
High Dim Statistics Workshop, ETH, Zurich, March, 2014
ITA, UC-San Diego, Feb, 2014
MSR, Redmond, Jan., 2014
Joint Seminar of Statistics and Biostatistics, UW, Seattle, Jan., 2014
Seminar at Statistical Science Center, Peking Univ, Dec. 2013
Joint Seminar of Statistics and Computer Science, UIUC, Dec., 2013
Simons Institute Workshop on Networks, Nov, 2013
Seminar, Biostatistics Department, Johns Hopkins University, Oct., 2013
Opening Workshop, SAMSI program on low dimensional structures in high dimension, Sept., 2013
Invited Session, JSM, Montreal Aug., 2013
Research Seminar, Microsoft Research Asia, Beijing, May, 2013
Young Researcher Statistics Conference, Beijing, May, 2013
Beijing Statistics Forum, Beijing, May, 2013
Statistics Seminar, CREST-ENSAE, Paris, March, 2013
Statistics Seminar, University of Padova, Italy, March, 2013
Statistics Seminar, Pau Fabra University, Barcelona, March, 2013
DARPA Worshop on big data, DC, March, 2013
Statistics Seminar, Columbia University, New York City, Jan., 2013
Simons Foundation roundtable on big data, New York City, Jan., 2013
Statistics Seminar, Department of Mathematics, Univ. of Oslo, Dec., 2012
Seminar, Department of Computer Science, University College London, Nov., 2012
The Oxford-Stanford Workshop on Big Data, Oxford Univ., Nov., 2012
Statistics Seminar, Center for Mathematical Sciences, Cambridge Univ., Nov., 2012
Research Seminar, ASA-LA Chapter, Oct., 2012
Statistics Seminar, Harvard University, Oct., 2012
Workshop on Massive Data, SAMSI, September, 2012
Joint Statistical Meetings (JSM), San Diego, Aug, 2012
New researcher meeting, JSM, San Diego, Aug., 2012
Young Researchers' Conference, the 8th World Congress of Probability and Statistics, Istanbul, July, 2012

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper	
02/16/2016	32.00	Julien Mairal, Bin Yu. Complexity Analysis of the Lasso Regularization Path, Proceedings of the 29th International Conferenceon Machine Learning. 26-JUN-12, . : ,
10/18/2012	8.00	Garvesh Raskutti, Martin J. Wainwright, Bin Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule, 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 27-SEP-11, Monticello, IL, USA. : ,
10/19/2012	14.00	Yanfeng Gu, Shizhe Wang, Tao Shi, Yinghui Lu, Eugene E. Clothiaux, Bin Yu. Multiple-kernel learning-based unmixing algorithm for estimation of cloud fractions with MODIS and CLOUDSAT data, IEEE International Geoscience and Remote Sensing Symposium. 22-JUL-12, . : ,

TOTAL: **3**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
02/11/2016	28.00	Siqi Wu, Antony Joseph, Ann S. Hammonds, Susan E. Celniker, Bin Yu, Erwin Frise. Stability driven nonnegative matrix factorization to interpret spatial gene expression and build localgene networks, Proceedings of the National Academy of Sciences of the United States of America (01 2016)
02/11/2016	27.00	Karl Rohe, Tai Qin, Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities, Proceedings of the National Academy of Sciences of the United States of America (01 2016)
02/16/2016	44.00	Siqi Wu, Bin Yu . Local identifiability of l1-minimization dictionary learning: asufficient and almost necessary condition, ArXiv:1505.04363v1 (05 2015)
07/31/2013	19.00	Bin Yu . Stability, Bernoulli (12 2012)
07/31/2013	15.00	Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of faithfulness assumption in causal inference, Annals of Statistics (08 2012)
07/31/2013	16.00	Yangbo He, Jinzhu Jia, Bin Yu. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs, Annals of Statistics (to appear) (03 2013)
10/18/2012	9.00	J. Mairal, B. Yu. Complexity analysis of the Lasso regularization path, (05 2012)
10/19/2012	10.00	B. Yu, J. Mairal. Supervised feature selection in graphs with path coding penalties and netwrok flows, (04 2012)
10/19/2012	13.00	Y. Wang, X. Jiang, B. Yu, M. Jiang. A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data, (07 2012)
10/19/2012	12.00	Y. Benjamini, B. Yu. The shuffle estimator for explainable variance in FMRI experiments., (07 2012)
10/19/2012	11.00	L. Miratrix, J. Sekhon, B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments, (09 2011)
11/06/2014	22.00	Antony Joseph, Bin Yu. Impact of regularization on Spectral Clustering, http://arxiv.org/abs/1312.1733v2 (12 2013)
11/06/2014	23.00	Sivaraman Balakrishnan, Martin J. Wainwright, Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis, http://arxiv.org/abs/1408.2156 (08 2014)
12/22/2011	1.00	G. Raskutti, M. Wainwrigt, B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming., Technical Report 795, Statistics Department, UC Berkeley. (12 2011)

12/22/2011	2.00	S. Negahban, P. Ravikumar, M. Wainwrigt, B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers., Statistical Science (accepted with minor revision) (10 2010)
12/22/2011	4.00	G. Rocha, B. Yu, S. N. Pakzad. Distributed modal identification by regularized auto regressive models., International Journal of Systems Science (to appear) (12 2011)
12/22/2011	5.00	V. Q. Vu, P. Ravikumar, T. Naselaris, K. N. Kay, J. L. Gallant, B. Yu. Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models., Annals of Applied Statistics (accepted) (04 2011)
12/22/2011	7.00	K. Rohe, S. Chatterjee, B. Yu. Spectral clustering and the high-dimensional Stochastic Block Model., Annals of Statistics (to appear) (12 2010)
TOTAL:	**18**	

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

STEM WOMEN Award, Nancy Skinner's Office, CA, 2014
Member, National Academy of Sciences, 2014

Thought-Leader Professor of Data Science, www.mastersindatascience.org, 2014
President, IMS, 2013-2014
Woodroofe Lecturer, Statistics Department, University of Michigan, Sept., 2013
Fellow, American Academy of Arts and Sciences, 2013
Distinguished Achievement Award, International Chinese Statistical Association, 2013
Inaugural Pao-Lu Hsu Award, International Chinese Statistical Association, 2012
President-Elect, IMS (Institute of Mathematical Statistics), 2012-2013
Tukey Memorial Lecturer in Statistics, the 8th World Congress in Probability and Statistics, the quadrennial joint conference of the Bernoulli Society and IMS, 2012

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Hongwei Li	0.05	
Shijing Yao	0.05	
FTE Equivalent:	**0.10**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Hanzhong Liu	0.75
FTE Equivalent:	**0.75**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Bin Yu	0.33	Yes
FTE Equivalent:	**0.33**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

Research discussions with researchers at Xerox European Research Center
Research discussions with researchers at LBNL
Consulting discussions with eBay researchers on A/B testing

Scientific progress and accomplishments

Under the ARO grant support (W911NF-11-1-0114), Bin Yu and her collaborators have conducted several research projects ranging from high dimensional statistical machine learning to methodology development motivated by interdisciplinary research in areas including systems biology, neuroscience, remote sensing, document summarization, and social networks.

1 Sparse and structured networks

The major goals of this project are to develop and implement algorithms based on high dimensional statistics theory, especially network theory, with the ultimate goal of extracting useful information from high dimensional data that arise from various fields of science and engineering.

Science and engineering abounds with different types of networks. Examples include social networks such as FaceBook and Twitter, networks of genes and proteins in molecular biology, network models for economic and market dynamics, neural networks in brain imaging, networks of disease transmission in epidemiology, and information networks in law enforcement. In the real-world, the structure of the underlying network is not known, but instead one observes samples of the network behavior (e.g., packet counts in a computer network; instances of infection at given time instances of an epidemic; emails or text messages sent among a group of people). Since the network data are complex, noisy and/or high-dimensional, it is challenging to infer the network structure. Developing methods for solving this network inference problem will have a broad range of applications. Examples include inferring brain connectivity and disease etiology in neuroimaging studies, detecting terrorist cells in social networks, monitoring intrusions in computer networks, and understanding the basis of gene-protein interactions in systems biology.

Spectral clustering is one of the popular techniques to identify communities (or clusters) in large network. The stochastic blockmodel is a social network model with well-defined communities; each node is a member of one community. In paper [21], we provide rigorous statistical analysis to the study of community detection by assessing how well spectral clustering can estimate the clusters in the Stochastic Blockmodel. Our results are the first clustering results that allow the number of clusters in the model to grow with the number of nodes, hence the name high-dimensional. In paper [8], we study the impact of regularization on spectral clustering and attempt to quantify the obtained improvement. We study in paper [7] the performance of spectral clustering in recovering the latent labels of i.i.d. samples from a finite mixture of nonparametric distributions. We provide a novel and useful characterization of the principal eigenspace of the population-level normalized Laplacian operator and establish a certain geometric property of nonparametric mixtures: embedded samples from different components are approximately orthogonal with high probability.

Aymmetric and undirected relationships are common assumptions in the clustering literature. However, the vast majority of relationships are asymmetric or directed. For example, in the gene regulatory network, one gene drives the transcription of the other gene. In the power grid network, electricity flows from one node to the other. In a communication network, one node initiate the conversation. In other examples, it might be more easy to observe the relationship without direction, but the direction remains of fundamental importance. For example, in a social network, a business searching for "trend leaders" wants to know the direction of influence in relationships. It is an interesting and important question to identify the clustering asymmetries in directed graphs. In paper [2], we propose a novel spectral co-clustering algorithm called DI-SIM for asymmetry discovery and directional clustering. A new Stochastic co-Block model is introduced to show favorable properties of DI-SIM. To accommodate sparse graphs and highly heterogeneous degrees within clusters, DI-SIM uses the regularized graph Laplacian and projection procedure. We apply a node-wise asymmetry score and DI-SIM to analyze the clustering asymmetries in the networks of Enron emails, political blogs, and the chemical connectome. In each example, a subset of nodes have clustering asymmetries; these nodes send edges to one cluster, but receive edges from another cluster. Such nodes yield insightful information (e.g. communication bottlenecks) about directed networks, but are missed if the analysis ignores

edge direction.

2 Systems biology

Yu has advanced the project with Dr. Frise et al on systems biology. Gene-gene interaction is at the heart of understanding regulatory pathways of organ formation and developmental disorders. Spatial gene co-occurrence information has been shown to be extremely useful in suggesting possible gene-gene interaction. The abundance of spatial gene expression data in recent years opens up an exciting new venue for reconstructing gene regulatory networks. However, due to the complexity of spatial gene expression and the noisy nature of the data acquiring process, extracting meaningful information from these data remains a challenge. In paper [1], we propose StaNMF method that combines a fast and scalable implementation of Non-negative Matrix Factorization (NMF) with a new stability-based criterion. StaNMF learns from a spatial gene expression data a set of data-driven basis, called Principal Patterns (PPs). As an example, using the spatial gene expression images of early stage embryonic Drosophila melanogaster, we demonstrate that the 21 learned PPs correspond to 21 localized pre-organ regions. The PPs provide a concise yet biologically interpretable representation, comparable to the well-established Drosophila fate map and serving as an alternative to human annotations. Based on the PPs, we construct spatially local correlation networks for all patterned transcription factors during early Drosophila development. With a two-tailed 2.5% cut-off, the constructed networks are consistent with 10 out of 12 links in the well-studied gap-gene network with six major gap genes. The very promising performance of PPs with the Drosophila data suggests StaNMF as a standard decomposition approach to examine complex and noisy gene expression data.

Our local network analysis recommends five uncharacterized genes as possible new candidates for the gap gene networks. Dr. Frises group and his collaborators have been working on CRISPR experiment to knock out each of the five candidate genes as experimental verification. So far, we learned that one of the genes are not viable, i.e. the fruit fly dies after the knock-out of the gene. Further examination of the ftz stained embryos indicates that the lack of the gene might lead to an elongated head and a wider gap between the first and the second segmentation stripes. Together with three students of mine, we are currently performing cell counting analysis in an effort to provide numerical evidence of our visual inspection conclusion.

Given the success of our approach for spatial gene expression analysis for early stage fruit fly embryos, we are in a process to extend it to model later stage gene expression. Due to the formation of internal organs of the embryo, registering the embryo unto a standard template for cross-individual comparison can be challenging. We build an organ classification and registration model that modifies state of the art computer vision algorithms to produce mid-level image features well suited to bio-imaging tasks. By combining our classification model with non-negative matrix factorization, we produce parts-based representations of spatial gene expression in various organ systems. Our PPs of gene expression are interpretable, low dimensional representations of the data that serve as a late stage analogue to the Drosophila fate map.

We have put large effort into automatization of our techniques for the benefits of system biology community. To facilitate automatic imaging, we contribute to designing a method that detects the most in-focus image as the microscope adjusts its focal plane. To scale up and speed up the computation for larger spatial gene expression data sets, we are working with Professor Andy Yao's group at Tsinghua University to create a general computation framework for biological data processing. Our system, called LSEMS, or Large Scale Experiment Management System, combines Gitlab for easy version control, MongoDB for storage of highly biological unstructured data and SPARK for distributed computing. The LSEMS framework allows biologists to submit a task on their personal laptops, triggering the remote machine to execute the task and distribute it to a large computer cluster. The results will be sent back to the user-end once the computation is completed. The initial testing reports that the system is much more efficient than the old one which uses a single machine. We are now building intuitive and easy-to-use GUI (graphical-user interface) to make the system more accessible to general biologists.

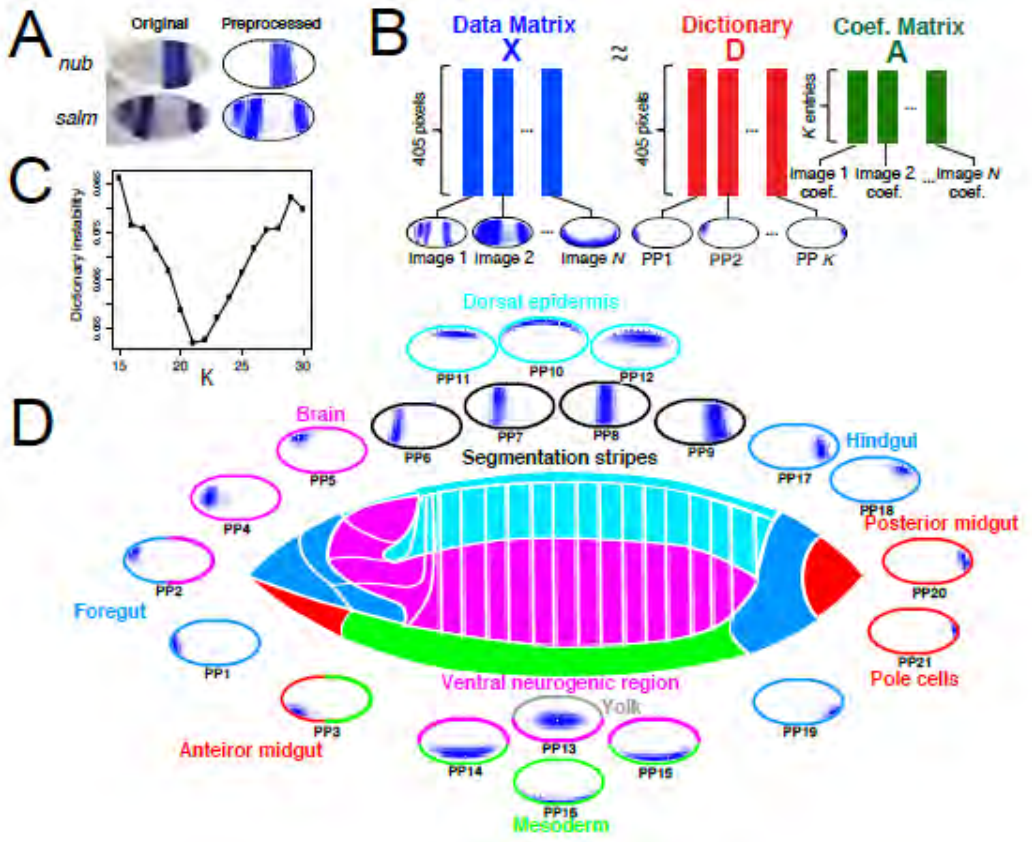

Figure 1: Learning principal patterns (PP) by staNMF from spatial gene expression patterns. (A) Expression patterns of two genes, nub and salm, in Drosophila embryos. (B) For a given number K, NMF factorizes the nonnegative data matrix X, the columns of which are gene expression images, into the product of two nonnegative matrices: dictionary D, which contains the K PP, and coefficient matrix A, which contains the nonnegative coefficients of the images. (C) StaNMF identified $K = 21$ to be the optimal number of PP for $15 \leq K \leq 30$. (D) The Drosophila fate map (center), surrounded by the 21 PP learned by staNMF. The PP are arrayed according to the corresponding regions of the fate map.

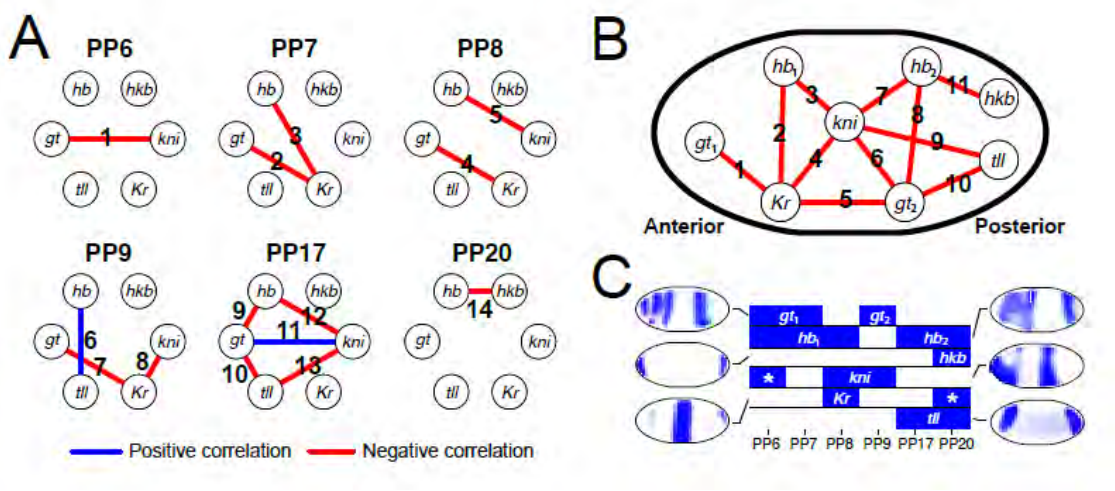

Figure 2: Modeling and validation of the Drosophila gap gene network with spatially local correlation networks (SLCN). (A) The SLCN for six gap genes. For each of the six gap-PP, shown is the sub-network of the SLCN that contains the six gap-genes. Links are numbered from 1 to 14. (B) The gap gene network diagram depicting repressive interactions of six genes. Links are numbered from 1 to 11 and multiple occurrence of the same gene are subscripted by numbers (e.g. hb1 and hb2). The directions of the interactions are not indicated. (C) Expression patterns of the six gap genes and their linearly ordered PP representation. For each gene, the regions depicted in blue are the gap-PP with sPP coefficient greater or equal to 0:1. The $*$ symbol indicates a region of gene expression with no match in (B).

3 Neuroscience

In computational neuroscience, it is important to estimate well the proportion of signal variance in the total variance of neural activity measurements. Paper [14] proposes a novel method to estimate the explainable variance in functional MRI (fMRI) brain activity measurements when there are strong correlations in the noise. Our shuffle estimator is nonparametric, unbiased, and built upon the random effect model reflecting the randomization in the fMRI data collection process. Motivated by collaborative research in neuroscience, papers [20, 15] answer questions under the Pearl causal inference framework, which is an alternative to the Neyman-Rubin framework. In particular, [20] proves analytical results to raise a red-flag on the commonly assumed faithfulness assumption. [15] proposes efficient MCMC algorithms to search through Markov equivalence classes of a causal graph. It was the first such algorithm that works fast enough for hundreds of nodes, admittedly under a sparsity condition.

Yu continues her collaborative work with the Gallant Lab on understanding visual pathway of primates by using sparse coding, invariant features and deep convolutional neural networks to build more accurate models of motion perception in the visual cortex. Our prior work has shown that image representations based the principles of sparse coding and nonlinear spatial pooling are empirically successful in explaining neural response in area V4 of the macaque visual cortex. Such models are able to discriminate between categories of image regions (such as foreground vs. background, texture vs. contour), while also generalizing across random realizations of object categories, due to their invariance to local deformation. These techniques have been adapted to modeling higher order visual areas such as area MT on two experimental datasets provided by the Gallant lab, in which the stimulus consisted of a series of short movie clips. This includes electrophysiological data from macaque area MT, as well as full visual cortex fMRI recordings of human subjects.

Deep convolutional neural networks are biologically inspired neural network based learning techniques. Recently, they have been the state-of-the-art methods for large-scale image recognition tasks in computer vision. We deploy the deep convolutional neural network features as early invariant features for modeling area V4. When combined with sparse linear modeling, we show that our deep convolutional feature based model outperforms the previous sparse coding based methods. Further more, our model not only has better prediction performance, but also leads to better interpretation, which could provide neuro-scientists clues about the receptive fields and orientation tuning preferences of individual neurons. We have also started to adapt deep convolutional neural networks techniques to model data collected from fMRI experiments. The main difficulty of modeling human neuronal activity via fMRI experiments is that extracting different meaningful features from video data that are general enough to model the dynamics different parts of the visual cortex area, ranging from the early sensory area such as LGN, V1 and V2, to higher order areas such as MT and IT. Our approach is to use a two stream deep convolutional neural networks, combining the spatial invariant features with temporal optical flow features. Our preliminary results show that the temporal optical flow features are not very good features when used alone for prediction, but when combined with spatial features, these temporal features largely improves prediction when compared to previous hard-crafted Gabor features based models.

3.1 Aerosol Optical Depth (AOD) retrieval

Yu has been collaborating with environmental scientists at JPL and Emory University to retrieval from NASA MISR remote sensing images aerosol index AOD for air pollution monitoring and management. Satellite-retrieved Aerosol Optical Depth (AOD) can potentially provide an effective way to complement the spatial coverage limitation of ground particulate air pollution monitoring network like AErosol RObotic NETwork (AERONET). Although the MISRs aerosol products lead to exciting research opportunities to study particle composition at a regional scale, its spatial resolution is too coarse for analyzing urban areas, where the air pollution has stronger spatial variations and can severely impact public health and the environment. Using NASA's novel multi-angle satellite sensor MISR, [16] develops a novel AOD retrieval algorithm with 4.4 km $\times$ 4.4 km resolution using Bayesian models and MCMC. [6] uses AERONET DRADON campaign data from 2011 in the Baltimore area to further validate our MCMC algorithm. We show that our MCMC algorithm substantially improves over the MISR operational algorithm both in terms of coverage and root-mean-square-error (RMSE).

4 Statistical guarantees of EM algorithm

The EM algorithm is a widely used tool in maximum-likelihood estimation in incomplete data problems. Existing theoretical work has focused on conditions under which the iterates or likelihood values converge, and the associated rate of convergence. Such guarantees do not distinguish whether the ultimate fixed point is a global or local optimum of the sample likelihood, nor its relation to the global optima of the idealized population likelihood (obtained in the limit of infinite data). In paper [3], we develop theoretical framework for quantifying when and how quickly EM-type iterates converge to a small neighborhood of a given global optimum of the population likelihood. For correctly specified models, such a characterization yields rigorous guarantees on the performance of two-stage estimators in which an initial pilot estimator is refined with iterations of the EM algorithm. Our analysis is divided into two parts: a treatment of the EM and gradient EM algorithms at the population level, followed by results that apply to these algorithms on a finite set of samples. We verify our conditions and give tight characterizations of the region of convergence for three canonical problems of interest: mixture of Gaussians, mixture of regressions, and linear regression with covariates missing completely at random.

5 Sparse modeling

Sparse models are necessary for model interpretability and computational efficiency in prediction. For example, expressing signals as sparse linear combinations of a dictionary basis has enjoyed great success in applications ranging from image denoising to audio compression. For certain data types such as natural image patches, predefined dictionaries like the wavelets are usually available. However, when a less-known data type is encountered, a new dictionary has to be designed for effective representations. Dictionary learning, or sparse coding, learns adaptively a dictionary from a set of training signals such that each signal has sparse representations under this dictionary. In paper [4], we study the theoretical properties of learning a dictionary from a set of N signals via l_1-minimization. We establish a sufficient and almost necessary condition for the reference dictionary to be locally identifiable, i.e. a local minimum of the expected l_1-norm objective function. With collaborators including students and postdocs, Yu has published five papers [10, 17, 18, 19, 11] on other topics of sparse modeling. The results provide insights on application of sparse classification to the problem of topic-specific summarization, when and why Lasso works under Poisson-like Heteroscedasticity, the complexity of Lasso solution path, minimax-optimal rates for sparse additive models over Kernel classes, and optimal data-dependent stopping rule of gradient descent for non-parametric regression.

6 Stability and inference

When data are perturbed (e.g. by subsampling), instability of results is common for big data, which are often high-dimensional. This instability begs a connection with Robust Statistics of Tukey and Huber. To bringing stability and hence interpretability and reproducibility to results of Lasso in high-dimension, [9] proposes an estimation stability (ES) metric to combine with the popular cross-validation (CV) for a dominant sparse modeling method Lasso (or ℓ_1- penalized Least Squares) that has been effective in our neuroscience work. For an image-fMRI data set from the Gallant Lab, we in fact improve interpretability substantially without losing prediction performance, relative to CV. [13] is an invited paper for a special issue of Bernoulli. It advocates for an enhanced emphasis on stability as a means to work towards reproducibility and promotes stability as a general statistical principle.

Inference and constructing confidence intervals for parameter estimation play important role in resulting interpretability findings. However, in high-dimensional setting, inference is challenging because the limiting distribution of estimators such as Lasso is complicated and hard to compute, which remains a barrier to widespread adoption of high-dimensional methodology in the sciences. In paper [12], we propose a valid inference procedure based on residual bootstrap after two-stage estimator Lasso+mLS (using Lasso to select a model and then using a modified version of Least Squares (mLS) refitting the coefficients in the selected model) and show consistency under suitable conditions. Compared with existing methods, such as debiasing, our method provides comparable results in terms of coverage probability and interval length, but our method is based on standard tools, the bootstrap and the Lasso, which is simple to implement and can be easily extended to models beyond linear regression.

7 Statistical analysis of algorithm leveraging

With rapid advances of information technology, massive datasets are collected by all fields of science, engineering, social science, business, and government. Useful or meaningful information is extracted from these data often through statistical means or model fitting, typically through regression models. These models are useful for predicting a response variable from p predictor variables or to describe relationships between predictor variables and a response variable. Given a set of n data units, in modern massive data sets, p and/or n can be large, in which case conventional algorithms face computational challenges. Subsampling of rows and/or columns of a data matrix has been employed traditionally as a heuristic to reduce the size of

large data sets. Recently, an innovative and effective sampling scheme based on using the empirical statistical leverage scores as a nonuniform importance sampling distribution has been proposed. The OLS based on such a subsample has been shown to give a good approximation to the OLS based on full data (when p is small and n is large), both in worst-case theory and in high-quality numerical implementation. The statistical properties of these algorithms are as of yet unexplored and are of interest for both fundamental and very practical reasons; and it is these properties that this project will address. One important question to be answered for using leverage subsampling for statistical estimation is: Under what conditions on p and n and the underlying model, the resulting leverage-OLS has good statistical properties such as a good mean-squared-error (MSE) when compared to the full-data OLS and other estimators, either in linear regression or non-linear regression models? Because of the noise properties in real data, it is challenging to answer this question.

In this project, we provide the first interpretation of algorithmic leveraging paradigm from a statistical analysis point of view. By performing a Taylor series analysis around the ordinary least-squares solution to approximate the subsampling estimators as linear combinations of random sampling matrices, we provide in paper [5] a simple yet effective framework to evaluate the statistical properties of algorithmic leveraging in the context of estimating parameters in a linear regression model with a fixed number of predictors. In particular, for several versions of leverage-based sampling, we derive results for the bias and variance, both conditional and unconditional on the observed data. We show that from the statistical perspective of bias and variance, neither leverage-based sampling nor uniform sampling dominates the other, which is particularly striking, given the well-known result that, from the algorithmic perspective of worst-case analysis, leverage-based sampling provides uniformly superior worst-case algorithmic results, when compared with uniform sampling. Based on these theoretical results, we propose and analyze two new leveraging algorithms: one constructs a smaller least squares problem with "shrinkage" leverage scores (SLEV), and the other solves a smaller and unweighted (or biased) least squares problem (LEVUNW). A detailed empirical evaluation of existing leverage-based methods as well as these two new methods is carried out on both synthetic and real data sets. The empirical results indicate that our theory is a good predictor of practical performance of existing and new leverage-based algorithms and that the new algorithms achieve improved performance. For example, with the same computation reduction as in the original algorithmic leveraging approach, our proposed SLEV typically leads to improved biases and variances both unconditionally and conditionally (on the observed data), and our proposed LEVUNW typically yields improved unconditional biases and variances.

References

[1] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu and E. Frise (2016). Stability driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences of the United States of America, accepted.*

[2] K. Rohe, T. Qin and B. Yu (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Submitted.*

[3] S. Balakrishnan, M. Wainwright and B. Yu (2015). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics, to appear.*

[4] S. Wu and B. Yu (2015). Local identifiability of l_1-minimization dictionary learning: a sufficient and almost necessary condition. *Submitted.*

[5] P. Ma, M. Mahoney, B. Yu (2015). A Statistical Perspective on Algorithmic Leveraging. *Journal of Machine Learning Research* **6**, 861-911.

[6] T. Moon, Y. Wang, Y. Liu and B. Yu (2015). Evaluation of a MISR-based high-resolution aerosol retrieval method using AERONET DRAGON campaign data. *IEEE Transactions on Geoscience and Remote Sensing*, **53**, 4328-4339.

[7] G. Schiebinger, M. J. Wainwright and B. Yu (2015). The geometry of kernelized spectral clustering. *The Annals of Statistics*, **43**, 819-846.

[8] A. Joseph and B. Yu (2015). The impact of regularization on spectral clustering. *Annals of Statistics, accepted*.

[9] C. Lim and B. Yu (2015). Estimation Stability with Cross Validation (ESCV). *Journal of Computational and Graphical Statistics (to appear)*.

[10] J. Jia, L. Miratrix, B. Yu, B. Gawalt, L. El Ghaoui, L. Barnesmoore, and S. Clavier (2014). Concise comparative summaries (CCS) of large text corpora with a human experiment. *The Annals of Applied Statistics* **8**, 499-529.

[11] G. Raskutti, M. J. Wainwright, and B. Yu (2014). Early stopping of gradient descent for non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, **15**, 335-366.

[12] H. Liu and B. Yu (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, **7**, 3124-3169.

[13] B. Yu (2013). Stability. *Bernoulli*, **19 (4)**, 1484-1500. (Invited paper for the Special Issue commemorating the 300th anniversary of the publication of Jakob Bernoullis Ars Conjectandi in 1712).

[14] Y. Benjamini, B. Yu (2013). The shuffle estimator for explainable variance in fMRI experiments. *Annals of Applied Statistics* **7**(4), 2007-2033.

[15] Y. He, J. Jia and B. Yu (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Annals of Statistics*, **41**(4), 1742-1779.

[16] Y. Wang, X. Jiang, B. Yu, M. Jiang (2013). A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data. *Journal of American Statistical Association*, **108**, 483-493.

[17] J. Jia, K. Rohe and B. Yu (2013). The Lasso under Poisson-like Heteroscedasticity. *Statistica Sinica*, **23**, 99-118.

[18] J. Mairal, B. Yu (2012). Complexity Analysis of the Lasso Regularization Path. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 353-360. .

[19] G. Raskutti, M. J. Wainwright, and B. Yu (2012). Minimax-Optimal Rates For Sparse Additive Models Over Kernel Classes Via Convex Programming. *Journal of Machine Learning Research*, **13**, 389-427.

[20] C. Uhler, G. Raskutti, and P. Buhlmann and B. Yu (2012). Geometry of faithfulness assumption in causal inference. *Annals of Statistics*, **41**, 436-463.

[21] K. Rohe, S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, **39**, 1878-1915.